UNSUPERVISED OUTLIER DETECTION

14th August 2019

Contents



[Executive Summary 3](#_Toc195692248)

[K Means Algorithm 4](#_Toc195692248)

[Classification and Outlier Detection 5](#_Toc195692248)

[Classification and Outlier Detection 6](#_Toc195692248)

[Optimal K value 7](#_Toc195692248)

[Dataset 8](#_Toc195692248)

[Normal Distributn plot 9](#_Toc195692248)

[Pre-processing 10](#_Toc195692248)

[Modelling and Prediction 12](#_Toc195692248)

[Conclusion 13](#_Toc195692248)

Executive Summary

My 3-month Industrial Internship Program work term was with EY India was involved in the area of e-Governance during my work term, all of which will be outlined in this report. This report will cover some background information on the projects I was involved in, as well as details on how the projects were developed.  
  
The project was on designing and implementing outlier detection for sales data using unsupervised learning.

I acquired many new technical skills throughout my work term. I acquired new knowledge in the area of Machine Learning. I used python to implement models for outlier detection. Most importantly, the work experience was very good which included good fellowship, cooperative teamwork and accepting responsibilities.

Although I spent a lot of time learning new things, I found that I was well trained in certain areas that helped me substantially in my projects. Many programming skills that I used in my projects, such as programming style and design, were ones that I had acquired during my studies in Computing Science.  
This report concludes with my overall impressions of my work experience as well as my opinion of the Industrial Internship Program in general

Outlier Detection using Unsupervised Learning

In machine learning, the detection of “not-normal” instances within datasets has always been of great interest. This process is commonly known as anomaly detection or outlier detection. An outlier in a pattern is dissimilar with rest of the pattern in a dataset. It has been used to detect and remove anomalous objects from data.

The main goal of this project is to learn through machine learning a model which can be used to pinpoint anomalies in the sales data. In this section the way such a model can be used will be explained as well as a general description of the process needed to obtain such a model. To obtain such a model the following steps need to be taken.

Types of Outliers:

**Type 1: Global Outliers (also called “Point Anomalies”):**  
A data point is considered a global outlier if its value is far outside the **entirety** of the data set in which it is found (similar to how “global variables” in a computer program can be accessed by any function in the program).

**Type 2: Contextual (Conditional) Outliers:**  
A data point is considered a contextual outlier if its value significantly deviates from the rest of the data points in the same context. Note that this means that same value may not be considered an outlier if it occurred in a different context. If we limit our discussion to [time series data](https://www.anodot.com/blog/closer-look-time-series-anomaly-detection/), the “context” is almost always temporal, because time series data are records of a specific quantity over time. It’s no surprise then that contextual outliers are common in time series data.

**Type 3: Collective Outliers:**  
A subset of data points within a data set is considered anomalous if those values as a collection deviate significantly from the entire data set, but the values of the individual data points are not themselves anomalous in either a contextual or global sense. In time series data, one way this can manifest is as normal peaks and valleys occurring outside of a time frame when that seasonal sequence is normal or as a combination of time series that is in an outlier state as a group.

K Means Algorithm

**K means** algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way k means algorithm works is as follows:

* Specify number of clusters K.
* Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
* Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.
* Compute the sum of the squared distance between data points and all centroids.
* Assign each data point to the closest cluster (centroid).
* Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Classification and Outlier Detection

The K-means clustering process results in cluster centroids. For the purpose of anomaly detection, we deploy two distance based methods – classification and outlier detection – that both use the K-means clustering results and that can be applied individually or in a combined way.

* Classification. The distances to the cluster centroids of the corresponding traffic class are calculated using the weighted Euclidean distance function. An object is classified as normal if it is closer to the normal cluster centroid than to the anomalous one, and vice versa. This distance-based classification allows detecting known kinds of anomalies, i.e. anomalous traffic with similar characteristics as in the training datasets.
* Outlier detection. An outlier is an object that differs from most other objects significantly. Therefore it can be considered as an anomaly. For outlier detection, only the distance to the appropriate centroid of the normal cluster is calculated. If the distance between an object and the centroid is larger than a predefined threshold dmax, the object is treated as an outlier.

Combined classification and outlier detection and anomaly. In contrast to the classification method, outlier detection does not make use of the anomalous cluster centroid, i.e. it may be less accurate in detecting known kinds of anomalies. On the other hand, it allows detecting new anomalies that do not appear in the training datasets. Combined classification and outlier detection. Classification and outlier detection can be used in a combined way in order to overcome the limitations of each individual method. If the two methods are applied simultaneously, an object is treated as an anomaly if it is closer to the anomalous cluster centroid than to the normal one, or if its distance to the normal cluster centroid is larger than the predefined threshold.

Optimal K Value

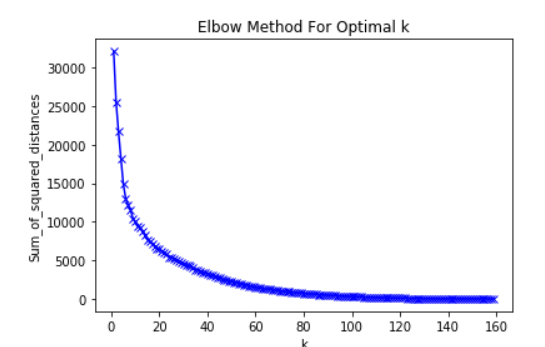
A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered.

The **Elbow method** is a method of interpretation and validation of consistency within [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) designed to help finding the [appropriate number of clusters in a dataset](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set).

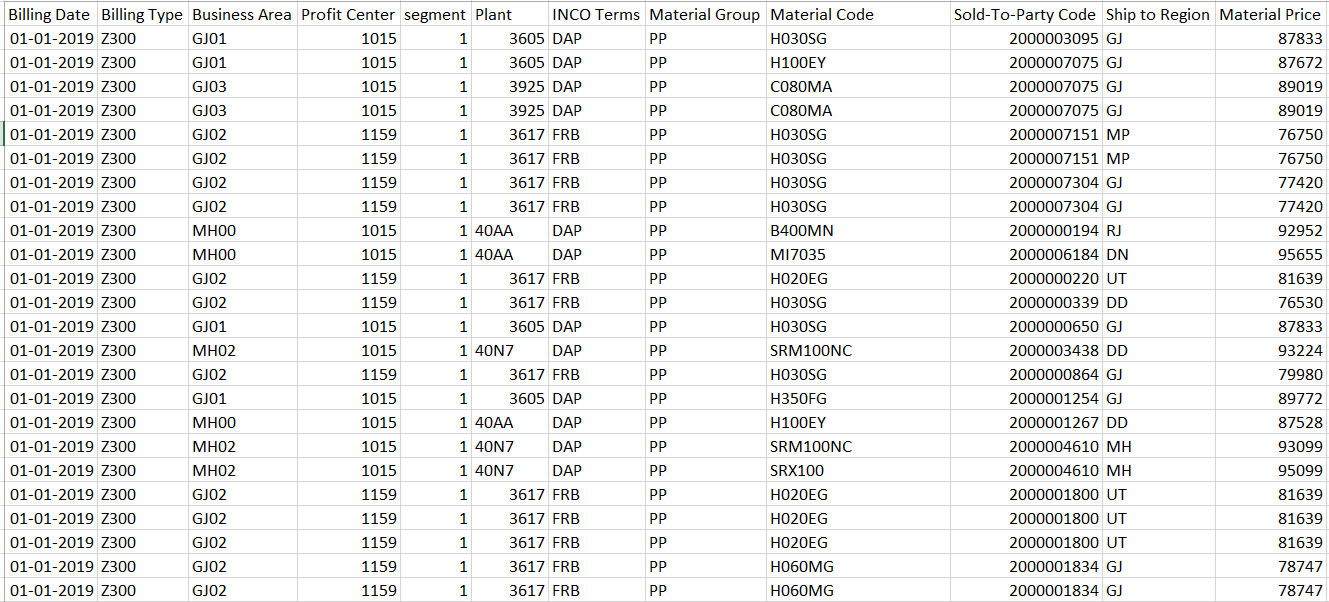
First of all, compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid.

If you plot k against the SSE, you will see that *the error decreases as K gets larger*; this is because when the number of clusters increases, they should be smaller, so distortion is also smaller. The idea of the elbow method is to choose the k at which the SSE decreases abruptly.

We iterate the values of k from 1 to 200 and calculate the values of distortions for each value of k and calculate the distortion for each value of k in the given range. When K increases, the centroids are closer to the clusters centroids. The improvements will decline, at some point rapidly, creating the elbow shape .That point is the optimal value for K. In the image above, K=150.

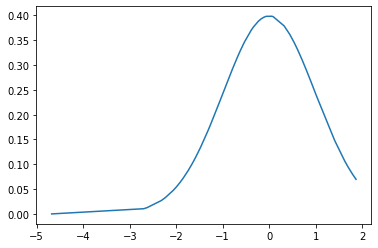


The Dataset



15000 X 12

Normal Plot for Material Price



This shows that the Normal Plot is left skewed.

Pre-processing the Data

* Only selective columns were considered for modelling. They were:
  + Billing Date : Date type
  + Billing Type : Category
  + Business Area : Category
  + Profit Center : Category
  + segment : Category
  + Sales Office : Category
  + Plant : Category
  + INCO Terms : Category
  + Material Group : Category
  + Ship to Region : Category
  + Material Price : Float type
  + Material Code : Category
* Removed all the missing values from these feature variables.
* Extracted the month from the Date Column and adding a month to the feature variables.
* Since the dataset contained too many material codes, I took only the top material code having the highest records.
* Billing date and material code was deleted as they weren’t needed for the modelling

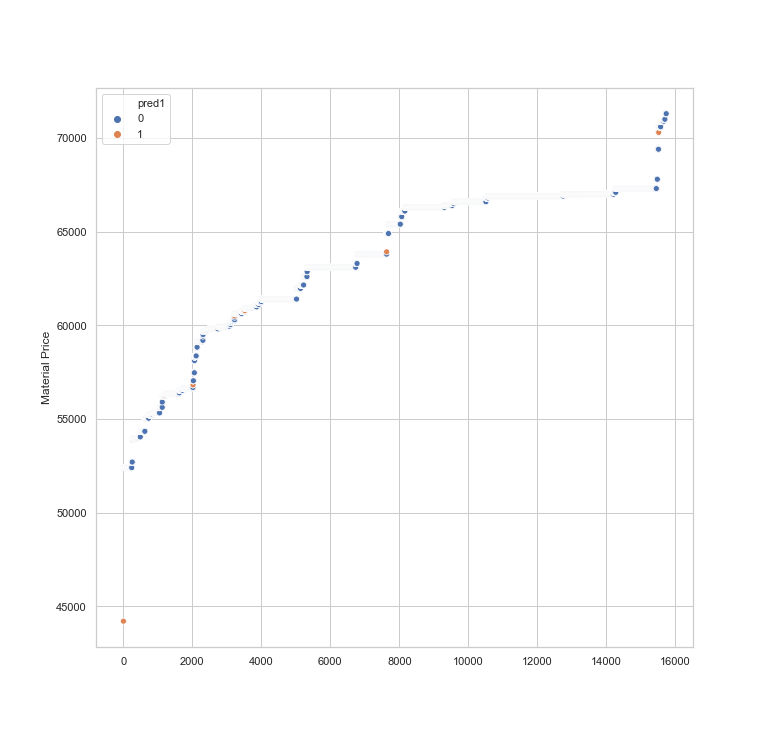
1. The features variables were encoded in one hot encoding and the material price was scaled

using standard scalar .

1. The feature variables are ready to be modelled for outlier detection.

Modelling and Prediction

The model was applied on 150 clusters and contamination of 0.005 which means 0.5% of the total dataset contains outliers.

The model provides a label for each dataset to distinguish between a normal point and anomaly.

Conclusion

I considered the problem of k-means with outliers and obtained a new algorithm. Our algorithm is simple, practical, and can be adapted to scale for large data; in addition, it has provable performance guarantees. I could close the gap between the desired number of outliers and the actual number of outliers found by the algorithm. I was successfully able to locate all the types of outliers global, contextual and collective.

Although it takes a lot of computational time to find the optimal K value using the inertia graph and also model it for large datasets.

**Offices of network firms of S.R. Batliboi & Affiliates**

**New Delhi**

4th Floor, Office 405

World Mark - 2, Asset No. 8

IGI Airport Hospitality District,

Aerocity

New Delhi – 110 037

**Noida**

7th Floor, Plot No. 2B

Tower 3, Sector – 126

Noida – 201 304

Gautam Budh Nagar,

Uttar Pradesh

**Pune**

C-401, Fourth Floor,

Panchshil Tech Park,

Yerwada,

Pune – 411 006

**Kochi**

9th Floor, Abad Nucleus

NH-49, Maradu

Kochi, Kerala – 682 304

**Kolkata**

22, Camac Street

3rd Floor, Block ‘B’,

Kolkata – 700 016

**Mumbai**

**Ruby**

12th Floor, The Ruby,

29 Senapati Bapat Marg

Dadar (West)

Mumbai – 400 028

**Goregaon**

5th Floor, Block-B-2,

Nirlon Knowledge

Park, Off W E Highway,

Goregaon (East),

Mumbai – 400 063

**Gurgaon**

**GVT, 2nd and 3rd Floor,**

Golf View Tower B,

Sector – 42, Sector Road,

Gurugram – 122 002,

Haryana

**Tower A, Building No. 8,**

1st Floor, DLF Cyber City,

Phase – II, Sector - 25,

Gurgaon – 122 002

Haryana

**Hyderabad**

Plot No.18, OVAL Office

ILABS Centre,

Software Units Layout,

Hitec City, Madhapur

Hyderabad – 500 081

**Ahmedabad**

Shivalik Ishan Building,

2nd Floor, Near CN Vidhyalaya,

Ambavadi,

Ahmedabad – 380 015

**Bengaluru**

12th Floor

“UB City” Canberra Block

No. 24, Vittal Mallya Road

Bengaluru – 560 001

**Chandigarh**

1st Floor, SCO: 166-167,

Madhya Marg, Sector 9-C

Chandigarh – 160 009

**Chennai**

6th Floor – “A” Block

Tidel Park

No. 4, Rajiv Gandhi Salai Taramani

Chennai – 600 113

**S R B C & CO LLP**

S R B C & CO LLP is a limited liability partnership firm of Chartered Accountants, registered with the Institute of Chartered Accountants of India, and having its registered office at 22 Camac Street, 3rd Floor, Block B, Kolkata – 700016.

© 2019 S R B C & CO LLP Published in India.

All Rights Reserved.

This publication contains information in summary form and is intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. Neither S R B C & CO LLP nor any of its fellow network firms can accept any responsibility for loss occasioned to any person acting or refraining from action as a result of any material in this publication.

This publication is meant for distribution internally, to our clients or distributable on specific requests only.

© 2008 S.R. Batliboi & Co

All Rights Reserved.

In line with S.R.Batliboi’s commitment to minimise its impact on the environment, this document has been printed on paper with a high recycled content.

Information in this publication is intended to provide only a general outline of the subjects covered. It should neither be regarded as comprehensive nor sufficient for making decisions, nor should it be used in place of professional advice. S.R. Batliboi & Co. accepts no responsibility for any loss arising from any action taken or not taken by anyone using this material.